**Algorithms for MLlib in Spark:**

|  |  |
| --- | --- |
| Spark-MLlib | |
| Feature Extractors | TF-IDF |
| Word2Vec |
| CountVectorizer |
| Feature Transformers | Tokenizer |
| StopWordsRemover |
| n-gram |
| Binarizer |
| PCA |
| PolynomialExpansion |
| Discrete Cosine Transform (DCT) |
| StringIndexer |
| IndexToString |
| OneHotEncoder |
| VectorIndexer |
| Normalizer |
| StandardScaler |
| MinMaxScaler |
| MaxAbsScaler |
| Bucketizer |
| ElementwiseProduct |
| SQLTransformer |
| VectorAssembler |
| QuantileDiscretizer |
| Feature Selectors | VectorSlicer |
| RFormula |
| ChiSqSelector |
| Classification | Logistic regression |
| Decision tree classifier |
| Random forest classifier |
| Gradient-boosted tree classifier |
| Multilayer perceptron classifier |
| One-vs-Rest classifier (a.k.a. One-vs-All) |
| Naive Bayes |
| Regression | Linear regression |
| Generalized linear regression( Gaussian/Binomial/Poisson/ Gamma) |
| Decision tree regression |
| Random forest regression |
| Gradient-boosted tree regression |
| Survival regression |
| Isotonic regression |
| Linear methods |  |
| Decision trees |  |
| Tree Ensembles | Random Forests |
| Gradient-Boosted Trees (GBTs) |
| K-means |  |
| Latent Dirichlet allocation (LDA) |  |
| Bisecting k-means |  |
| Gaussian Mixture Model (GMM) |  |
| Collaborative filtering | alternating least squares(ALS) |
| Model selection (a.k.a. hyperparameter tuning) |  |
| Cross-Validation |  |
| Train-Validation Split |  |

**参考文献：**

<http://spark.apache.org/docs/latest/ml-guide.html>

**[List of machine learning algorithms](https://en.wikipedia.org/wiki/List_of_machine_learning_algorithms):**

|  |  |  |
| --- | --- | --- |
| Types |  |  |
| Supervised learning | AODE |  |
| Artificial neural network |  |
| Bayesian statistics |  |
| Case-based reasoning |  |
| Gaussian process regression |  |
| Gene expression programming |  |
| Group method of data handling |  |
| Inductive logic programming |  |
| Instance-based learning |  |
| Lazy learning |  |
| Learning Automata |  |
| Learning Vector Quantization |  |
| Logistic Model Tree |  |
| Minimum message length |  |
| PCA |  |
| Ripple down rules |  |
| Symbolic machine learning |  |
| Support vector machines |  |
| Random Forests |  |
| Ensembles of classifiers(Bootstrap/Boosting) |  |
| Ordinal classification |  |
| Information fuzzy networks (IFN) |  |
| Conditional Random Field |  |
| ANOVA |  |
| Linear classifiers | Fisher's linear discriminant |
| Linear regression |
| Logistic regression |
| Multinomial logistic regression |
| Naive Bayes classifier |
| Perceptron |
| Support vector machines |
| Quadratic classifiers |  |
| k-nearest neighbor |  |
| Boosting |  |
| Decision trees |  |
| Bayesian networks |  |
| Hidden Markov models |  |
| Unsupervised learning | Expectation-maximization algorithm |  |
|  | Vector Quantization |  |
|  | Generative topographic map |  |
|  | Information bottleneck method |  |
|  | Artificial neural network | Self-organizing map |
|  | Association rule learning | Apriori algorithm |
|  |  | Eclat algorithm |
|  |  | FP-growth algorithm |
|  | Hierarchical clustering | Single-linkage clustering |
|  |  | Conceptual clustering |
|  | Cluster analysis | K-means algorithm |
|  |  | Fuzzy clustering |
|  |  | DBSCAN |
|  |  | OPTICS algorithm |
|  | Outlier Detection | Local Outlier Factor |
| Semi-supervised learning | Generative models |  |
|  | Low-density separation |  |
|  | Graph-based methods |  |
|  | Co-training |  |
| Reinforcement learning | Temporal difference learning |  |
|  | Q-learning |  |
|  | Learning Automata |  |
|  | SARSA |  |
| Deep learning |  |  |

**参考文献：**

<https://en.wikipedia.org/wiki/List_of_machine_learning_concepts>

**Machine Learning Appraoches:**

|  |  |
| --- | --- |
| Decision tree learning |  |
| Association rule learning | Apriori algorithm |
| Eclat algorithm |
| FP-growth algorithm |
| Artificial neural networks |  |
| Deep learning |  |
| Inductive logic programming |  |
| Support vector machines | Linear SVM |
| Nonlinear classification |
| Clustering | hierarchical clustering |
| DBSCAN |
| OPTICS |
| HCS clustering algorithm |
| Bayesian networks |  |
| Reinforcement learning |  |
| Representation learning |  |
| Similarity and metric learning |  |
| Sparse dictionary learning |  |
| Genetic algorithms |  |
| Rule-based machine learning |  |
| Learning classifier systems |  |

**参考文献：**

<https://en.wikipedia.org/wiki/Machine_learning>

**Spark和Scikit-learn机器学习算法对比：**

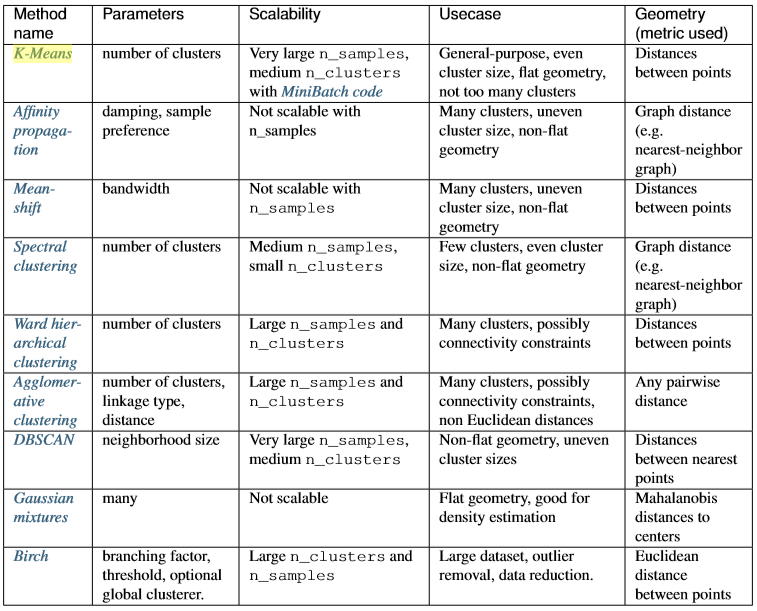
|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 算法 | Spark | scikit-learn |
| 分类 | KNN |  | 1 |
| Decision trees | 1 | 1 |
| 朴素贝叶斯 | 1 | 1 |
| Logistic regression | 1 | 1 |
| 线性SVM | 1 | 1 |
| Random Forests | 1 | 1 |
| 梯度增强树(GBRT) | 1 | 1 |
| SVC |  | 1 |
| Gaussian Process Classiﬁcation (GPC) |  | 1 |
| Neural network models |  | 1 |
| 回归 | LASSO | 1 | 1 |
| 线性最小二乘 Ordinary Least Squares | 1 | 1 |
| Ridge Regression | 1 | 1 |
| Elastic Net |  | 1 |
| Least Angle Regression(LARS) |  | 1 |
| Orthogonal Matching Pursuit (OMP) |  | 1 |
| Bayesian Regression |  | 1 |
| Automatic Relevance Determination随机抽样一致性回归(ARD) |  | 1 |
| Logistic regression |  | 1 |
| Passive Aggressive Algorithms |  | 1 |
| RANSAC: RANdom SAmple Consensus |  | 1 |
| Huber Regression |  | 1 |
| Kernel ridge regression |  | 1 |
| Support Vector Machines |  | 1 |
| SVR |  | 1 |
| Gaussian Process Regression (GPR) |  | 1 |
| 决策树 | 1 | 1 |
| 随机森林 | 1 | 1 |
| 梯度增强树 | 1 | 1 |
| 保序回归 | 1 | 1 |
| Neural network models |  | 1 |
| 其他有监督学习 | 典型相关分析CCA |  | 1 |
| 偏最小二乘PLS |  | 1 |
| 无监督学习（聚类） | k-means算法 | 1 | 1 |
| GMM（高斯混合模型） | 1 | 1 |
| PIC（快速迭代聚类） | 1 |  |
| LDA（隐式狄利克雷分布) | 1 | 1 |
| 二分k-means算法 | 1 |  |
| 流式k-means算法 | 1 |  |
| Afﬁnity Propagation |  | 1 |
| Mean Shift |  | 1 |
| Spectral clustering |  | 1 |
| Different label assignment strategies |  | 1 |
| Hierarchical clustering |  | 1 |
| 无监督学习（其他） | 关联分析Apriori |  |  |
| 频繁项集FP-growth |  |  |
| 无监督学习（降维） | PCA主成分分析 | 1 | 1 |
| SVD简化数据 | 1 | 1 |
| EVD | 1 |  |
| Isomap[PCA/(MDS的扩展] |  | 1 |
| Locally Linear Embedding |  | 1 |
| Hessian Eigenmapping |  | 1 |
| Spectral Embedding |  | 1 |
| DBSCAN |  | 1 |
| Birch |  | 1 |
| Ward |  | 1 |
| AgglomerativeClustering |  | 1 |
| Gaussian Mixtures |  | 1 |
| NMF/NNMF |  | 1 |
| LDA（Latent Dirichlet Allocation ） |  | 1 |
| 其他无监督学习 | Novelty and Outlier Detection:Isolation Forest |  | 1 |
| Density Estimation |  | 1 |
| Neural network models |  | 1 |
| 推荐 | Collaborative filtering ALS | 1 |  |
| 最优化方法 | 梯度下降 | 1 | 1 |
| L-BFGS | 1 | 1 |
| 非负最小二乘 | 1 |  |
| newton-cg |  | 1 |
| sag |  | 1 |

**参考文献：**

Machine Learning in Action, 机器学习实战, Peter Harrington.

scikit-learn user guide, *Release 0.18.1.*

**scikit-learn clustering：**



**参考文献：**

scikit-learn user guide, *Release 0.18.1.*